

Article

Deep-Learning-Based Accurate Identification of Warehouse Goods for Robot Picking Operations

Huwei Liu ^{1,2} , Li Zhou ¹, Junhui Zhao ¹, Fan Wang ¹, Jianglong Yang ^{1,2,*} , Kaibo Liang ¹  and Zhaochan Li ¹

¹ School of Information, Beijing Wuzi University, Beijing 101149, China; darion8@163.com (H.L.); zhouli@bwu.edu.cn (L.Z.); z19834405561@163.com (J.Z.); 13137521516@163.com (F.W.); liangkaibo2955@gmail.com (K.L.); darion6@163.com (Z.L.)

² School of Management and Engineering, Capital University of Economics and Business, Beijing 100070, China

* Correspondence: darion8@yeah.net

Abstract: In order to explore the application of robots in intelligent supply-chain and digital logistics, and to achieve efficient operation, energy conservation, and emission reduction in the field of warehousing and sorting, we conducted research in the field of unmanned sorting and automated warehousing. Under the guidance of the theory of sustainable development, the ESG (Environmental Social Governance) goals in the social aspect are realized through digital technology in the storage field. In the picking process of warehousing, efficient and accurate cargo identification is the premise to ensure the accuracy and timeliness of intelligent robot operation. According to the driving and grasping methods of different robot arms, the image recognition model of arbitrarily shaped objects is established by using a convolution neural network (CNN) on the basis of simulating a human hand grasping objects. The model updates the loss function value and global step size by exponential decay and moving average, realizes the identification and classification of goods, and obtains the running dynamics of the program in real time by using visual tools. In addition, combined with the different characteristics of the data set, such as shape, size, surface material, brittleness, weight, among others, different intelligent grab solutions are selected for different types of goods to realize the automatic picking of goods of any shape in the picking list. Through the application of intelligent item grabbing in the storage field, it lays a foundation for the construction of an intelligent supply-chain system, and provides a new research perspective for cooperative robots (COBOT) in the field of logistics warehousing.

Keywords: warehouse picking; neural network; goods identification; intelligent grabbing



Citation: Liu, H.; Zhou, L.; Zhao, J.; Wang, F.; Yang, J.; Liang, K.; Li, Z. Deep-Learning-Based Accurate Identification of Warehouse Goods for Robot Picking Operations. *Sustainability* **2022**, *14*, 7781. <https://doi.org/10.3390/su14137781>

Academic Editor: Giada La Scalia

Received: 7 June 2022

Accepted: 23 June 2022

Published: 26 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The demand of consumers is developing in the direction of diversification and personalization, so the accuracy and immediacy of orders become important factors to improve customer satisfaction. In order to make the operation of logistics storage system more efficient and flexible, and further improve the picking efficiency, many enterprises have begun to introduce intelligent picking robots. In the field of warehousing and logistics, there is a boom of “robots replacing humans”, and this trend will continue to develop. Jingdong Asia No.1 automated stereoscopic warehouse is in the leading position in China [1]. In its storage area of general merchandise, the storage boxes where the goods are located can leave the shelf quickly through the sorting equipment and be transported to the picking table through the automatic conveyor belt. The process of goods from the shelf to the picking table is fully automatic, but there are only a few air suction robot arms to pick the articles with regular shape in the picking table. In addition, there are still a large number of picking tasks that are completed manually through the Kanban system. Jingdong Asia No.1 is known as an unmanned automated warehouse. If goods can be fully grasped by robot arms on the picking table and packaging automation is realized, the real automated warehouse can be realized.

In the process of robot picking, the efficient identification of goods is one of the key links for correct picking. Especially in the complex logistics picking environment, we can use the neural network model of deep learning to identify the goods in the warehouse, reasonably classify all the identified goods, match different types of intelligent robot grasping methods for different types of goods, and further realize the intelligent and efficient warehouse picking operation. Taking some unmanned warehouses of large-scale e-commerce logistics as the application background, this paper uses a neural network to establish the goods identification model, and then uses the robot with vision to complete the identification of intelligent grasping methods according to the characteristics of different types of goods.

2. Related Works

2.1. Target Identification and Classification

Based on the idea of digital transformation, when the supply chain is combined with high-tech technologies such as big data, the decision-making will be more accurate and efficient [2]. From the perspective of economics, the idea of digital artificial intelligence is used to build a close relationship between high-tech and digital concepts [3]. In the main practical applications of computer recognition technology, one is target detection and the other is target classification, and as artificial intelligence computer vision, this approach is also closely related to digitization. In the aspect of target detection, some are based on the R-CNN framework and YOLO v3 algorithm to locate and identify the grid points and crops, and realize the positioning and position calibration within the spatial range [4,5]; some intelligent object surface detection systems based on the fast R-CNN algorithm can realize the surface detection and detect the location of geometrically complex products [6]. Further, there are goods detection and recognition based on multilevel CNN, deep CNN, and deep neural networks to realize the detection of unsafe articles in subway security inspection and operation, and improve the recognition accuracy [7–9]. Furthermore, the random maximum interval combination model constructed by a weak information part and R-CNN detection are used to realize the accurate recognition of objects [10,11]. Zhao et al. [12] proposed a multilevel feature pyramid network (MLFPN) to detect objects of different scales. Xie et al. [13] proposed a new dimension decomposition area recommendation network (DeRPN), which solved the imbalance problem of target loss calculation at different scales. Alom et al. [14] established an initial residual convolution neural network (IRRCN) for target recognition, which improved the recognition accuracy of the initial residual network with the same number of network parameters. In the aspect of target classification, some use CNN, self encoder, and depth neural networks to automatically classify remote sensing images to realize the recognition and classification of different buildings in the region [15–17], some use RPN and fast R-CNN to diagnose and classify medical images, and apply image classification to knee OA automatic diagnosis [18]. Additionally, there is gesture recognition classification based on CNN, and the recognition rate of eight types of gestures reaches 98.52% [19]. Hu et al. [20] proposed a gesture recognition system designed to control the flight of unmanned aerial vehicles (UAV), and introduced a data model that represents dynamic gesture sequences by converting 4D spatiotemporal data into 2D matrices and 1D arrays. Pigou et al. [21] explored a deep architecture for gesture recognition in video and proposed a new end-to-end trainable neural network that combines temporal convolution and two-way recursion. Fast R-CNN and an artificial neural network are also applied to automatic classification of fruits and beans, making detection and classification more accurate and rapid [22,23]. Kittinun et al. [24] proposed a solution to locate and classify rice grains in images, developed a deep-learning-based solution, and trained a masked-region-based convolutional neural network (R-CNN) to locate and classify images. In addition, the multilayer perceptron neural network, trained by a hybrid population-physics-based algorithm, the combination of particle swarm optimization and the gravitational search algorithm, was proposed for the classification of an encephalography data set [25]. Additionally, the recursive neural

network is also used to classify the image, and compared with the convolution neural network the computational cost is significantly reduced [26].

2.2. Robot Picking

In the field of logistics and warehousing, the application of computer identification technology can realize the rapid and accurate identification and classification of goods, provide a prerequisite for intelligent picking of goods, and improve the picking efficiency in the logistics field. According to the different driving structure and power source, robot arms are mainly divided into hydraulic drive, pneumatic drive, electric drive, and mechanical drive; the characteristics of the different types of robot arms are shown in Table 1. Additionally, there are many types of executive structures for common robot arms, such as grip-type, hook-support-type, spring-type and air-suction-type; the applications of various types of robot arms are shown in Table 2.

Table 1. Characteristics of robotic arms with various driving methods.

Driving Methods	Structural Features	Freedom of Installation and Control	Output Grip	Positioning Accuracy	Speed of Response	Reliability	Cost
Hydraulic	Complex	High	Strong	General	Fast	High	High
Pneumatic	Simple	High	Weak	General	Slow	Low	Low
Electric	Complex	Medium	Weak	Very high	Fast	Medium	Low
Mechanical	Medium	Low	Weak	Very high	Fast	Low	Low

Table 2. Situation of various robotic arms.

Crawl Method	Working Principle	Structural Features	Application
Hook bracket	The robot uses the actions of holding, hooking and holding to hold the goods.	The structure is simple, the driving requirements are low, and the horizontal or vertical conveying operation can be completed.	Suitable for large machinery and equipment.
Spring-loaded	The grasping action mainly depends on the force of the spring to clamp the goods.	No special drive is required.	Mostly used to grab small and light goods.
Grab-type	Grabs the goods with mechanical fingers.	According to the target product, different shapes and numbers of mechanical fingers are designed.	It is most common in industrial robots.
Air suction	The suction cup compresses the gas and generates adsorption force through the pressure difference.	Simple structure, easy to buy, easy to operate, and low requirements for goods positioning.	Wide range of applications, and especially when only one side of the product can be contacted, the air suction type is the best.

In addition, the robot's perception ability is usually used to recognize a given object and then grasp it. The key is how to control the end effector of the robot to ensure the safety and efficiency in the grasping process. Taking rough-processed construction and demolition waste as an example, its grasping strategy on the conveyor belt was discussed and its function of grasping and sorting on the recycling line was implemented with a robotic locating system established by mathematical models [27]. Zhang [28] proposed a robot sorting system based on computer vision, established a plane vision coordinate system by using a CCD image sensor, and solved the shortcomings of low efficiency and low precision of target recognition through the deep-learning convolution algorithm. Morrison et al. [29] proposed a real-time generation method of robot grasping, and established the grasping convolution neural network (GG-CNN) to directly generate the grasping posture from the pixel-based depth image. Arapi et al. [30] proposed a framework for soft robot arms grasping fault prediction based on deep learning. By training the network, the problem of fault prediction of flexible robot arms before grasping was solved and the occurrence of the fault was predicted. Fang et al. [31] proposed a task-oriented grasping network (TOG-NET),

which optimizes the grasping action of the robot with the aid of tools, and jointly optimizes the task-oriented grasping model and the operation strategy of corresponding tools through large-scale self-monitoring training. Jiang et al. [32] proposed a nontextured-planar-object vision-guided robot picking system based on a depth image, and used a deep CNN (DCNN) model to train the depth image generated in the physical simulator, and directly predicted the grasping point. De Conink et al. [33] proposed a system for robots to learn how to grasp arbitrary objects through visual input, and used an intelligent grasp controller based on CNN to realize the selection and placement tasks, and to adapt to the environment of moving objects. Wang [34] proposed an end-to-end deep-learning method based on the point cloud, which fused the point cloud and class number into a point class vector, and used multiple point network class networks to estimate the grasping posture of the robot. Yang et al. [35] aimed at the problem of positioning accuracy of a stereovision-guided robot, TSAI hand-eye calibration method was used to calculate the relative position and posture of the robot end relative to the left camera. The experiment showed that the method could accurately guide the robot to accurately locate the target point.

At present, the research on intelligent robot picking mainly focuses on picking path optimization, order-by-order delivery, loading and unloading, and intelligent scheduling of multiple robots; less research is based on neural network or machine learning in object recognition and intelligent grabbing. Neural networks have been widely used in many fields and have achieved success, but further verification is needed in logistics picking. Deep-learning algorithms based on neural networks can achieve more efficient image recognition, which is necessary to identify goods in logistics storage image information. From the perspective of the development of warehouse automation, the movement of shelves or cargo boxes has reached a high level of intelligence, but many of the picking operations are still manual. Grabbing of goods is still a weak link in the entire picking process. Therefore, it is necessary to study how to use robots to select according to the type of goods, and to improve efficiency. More importantly, in the field of intelligent capture of e-commerce warehouse picking, there is a lack of theoretical methods on generalized sustainable development. By completing the intelligent capture task and realizing a higher level of personalization of e-commerce warehousing, we can not only achieve green development in energy conservation and emission reduction, but also improve the digitalization and flexibility of warehousing picking operations, so as to continuously improve the operation efficiency of e-commerce warehousing. Exploring the security cooperation between human and robot in the management of intelligent supply-chain and digital logistics is helpful for implementing digital technology in achieving ESG (Environmental Social Governance) goals in society.

In this study, crawler technology was used to obtain pictures of e-commerce warehouse products, and the categories were divided to match the grasping methods of different types of robots. According to the deep-learning algorithm, we established a visual recognition model based on a neural network, which was used to identify the type of items to be grabbed in the logistics storage environment, and complete the task of matching the type of items with the type of manipulator. The training and testing of the model were visualized, which verified the validity of the model, provided new ideas for the application research of sorting robots in logistics and warehousing activities, and broadened the research scope of intelligent robots in warehousing and sorting logistics. Through the neural network and robot visual recognition technology in deep learning, the related research on robot intelligent grabbing in logistics and warehousing, to some extent, has been promoted. In the case of COBOT (cooperative robot), the original human-computer cooperative picking and grabbing link is improved to use a mechanical arm with a visual recognition mechanism for intelligent grabbing, so as to replace manual operation and obtain a practical application for the picking link of logistics warehousing. On this basis, we can build an intelligent supply chain that includes the safe cooperation between workers and COBOT. Therefore, this paper puts forward a new research direction to achieve the ESG goal by improving the intelligent grasping function of the logistics warehousing robot.

3. Proposed Approach

3.1. Acquisition and Processing of Picture Data

In this study, a web crawler was used to obtain picture information on open websites of three large domestic e-commerce companies in China, which information was then preprocessed. The products identified were mainly common household daily department store items, which were mainly divided into the following eight categories, as shown in Table 3: facial toiletries, freezing refrigeration, stationery, drinks and beverages, laundry supplies, puffed food, paper products, miscellaneous department store items. These pictures number 33,932, and a part of each type of picture is extracted for display in Figure 1.

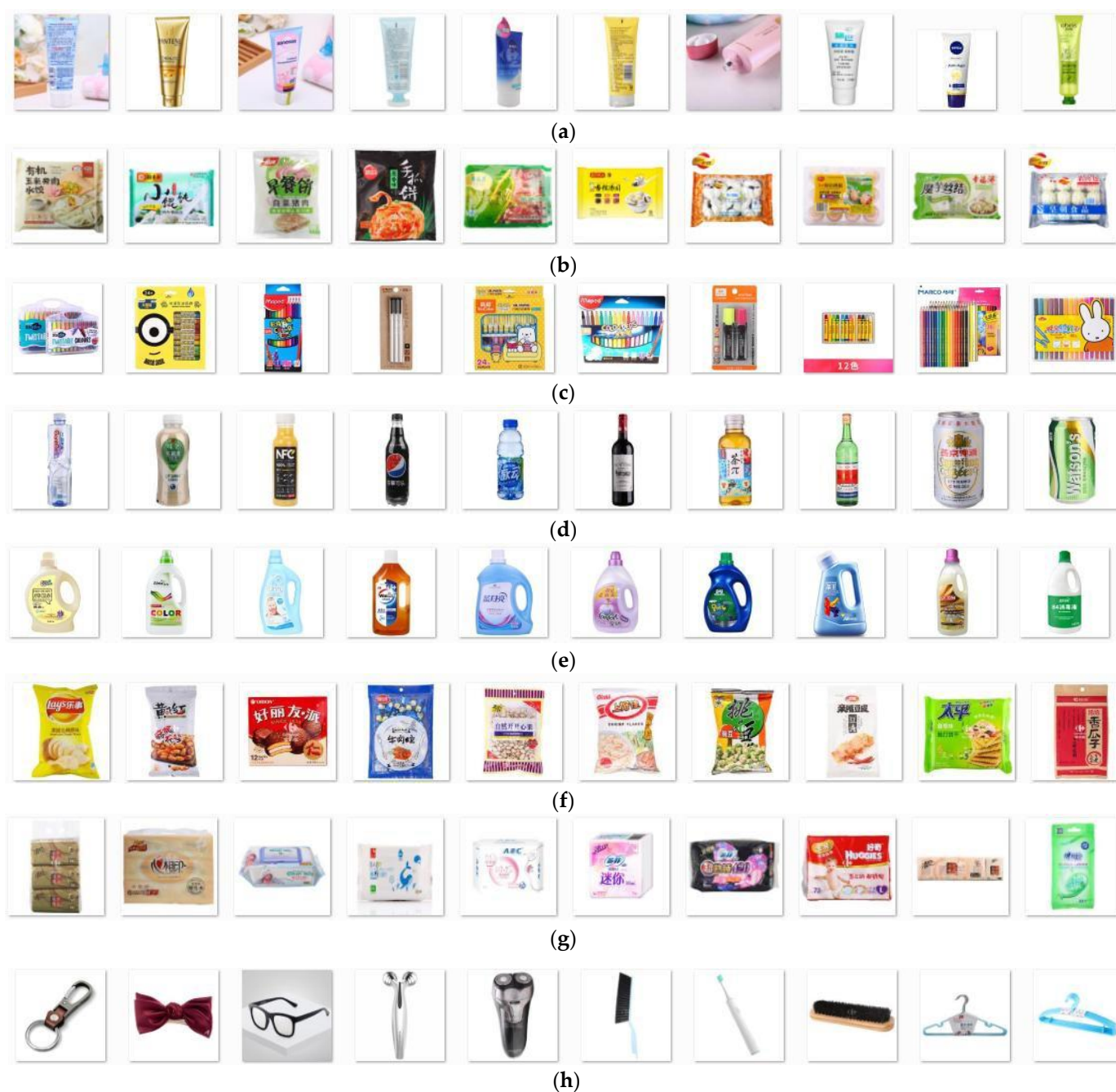


Figure 1. Part of each type of picture extracted: (a) facial toiletries; (b) freezing refrigeration; (c) stationery; (d) drinks and beverages; (e) laundry supplies; (f) puffed food; (g) paper products; (h) miscellaneous department store items.

Table 3. Data of common household department stores.

	Classification Name	Content	Number of Images	Example
1	Facial toiletries	Cleansing cream, sun cream, eye cream, lotion, etc.	2537	Figure 1a
2	Freezing and refrigerated	Noodles, vegetarian food, dumplings, sausages, pot vegetables, fast food, etc.	5626	Figure 1b
3	Stationery	Crayons, pencils, crayons, pastels, etc.	5635	Figure 1c
4	Drinks and beverages	Tea drinks, functional drinks, fruit juices, carbonated drinks, spirits, wines, fruit wines, beer, drinking water, etc.	1784	Figure 1d
5	Laundry supplies	Laundry liquid, laundry disinfectant, etc.	2407	Figure 1e
6	Puffed food	Biscuits, pastries, rice crackers, roasted nuts, snacks, etc.	6906	Figure 1f
7	Paper products	Paper, wipes, sanitary napkins and pads, etc.	5425	Figure 1g
8	Miscellaneous department store items	Hangers, shoe brushes, toothbrushes, hooks, keychains, table lamps, etc.	3216	Figure 1h

Because the original data obtained by the web crawler often has some abnormal conditions that may affect the training and testing results of the model, the data need to be cleaned before use, for example: delete invalid data in the original data set, smooth out noisy data, delete data that have nothing to do with the subject, and manage missing values, outliers, etc. Through data cleaning, 30,000 pieces of data that contain 30,000 goods and meet the requirements were screened. In addition, to further improve the recognition effect of the model and prevent the occurrence of overfitting, the image data of the product were expanded by random rotation. In order to meet the model's requirements for data input consistency, the input picture data were processed uniformly, the color picture size changed to $3 \times 32 \times 32$ pixels, and the product picture data set file divided into training and test data sets to train and test the model.

3.2. Grasping Solutions of Goods

According to the analysis of the characteristics of the robot arms' power and grasp method, combined with the actual situation of the target goods and the demand for the grasp of robot arms, this study selected the electric-driven, pneumatic-driven, hydraulic-driven grabbing, and air-suctioning robot arms to complete the grab work of various types of goods. For grabbing robot arms, rubber and foam materials were used to wrap the finger surface, so that the finger surface had a certain softness, and robot fingers could produce greater friction when grasping, and would not scratch the surface of the goods. Air-suctioning robot arms were used to grasp the goods with smooth and clean outer packages. When the robot arm was adsorbing, there was no air gap on the contact surface, which could achieve a good grasping effect. Considering the factors such as the type of goods, material, area, volume and weight, a more targeted and practical solution was designed and combined with the driving and grasping methods, as shown in Table 4.

Table 4. Specific grabbing schemes for different types of goods.

Cargo Variety	Specification	Features	Drive Methods	Fetching Solution
Facial toiletries	100~150 g	Packaging of plastic cylinder.	Pneumatic	Gripping of variable diameter and one-hand pneumatically driven.
Freezing and refrigerated	200~1000 g	Some quotients are soft.	Electric	Flexible hands holding.
Stationery	40~100 g	Packaging of plastic strip.	Pneumatic	Pneumatic grip with both hands.
Drinks and beverages	235~2000 mL	Glass and plastic bottles.	Pneumatic	Three-jaw chuck with suction cup.
Laundry supplies	110~4260 g	Plastics pot.	Hydraulic	Hydraulically driven with both hands.
Puffed food	16~70 g	Lightweight vacuum bag.	Pneumatic	Bionic software robot with suction cup for grasping.
Paper products	300~400 g	Bag is the smallest unit and resistant to crushing.	Electric	Electric-powered and two-handed telescopic splint for gripping.
Miscellaneous department store items	—	No obvious regularity in shape.	Electric	Grab by electric-powered bionic robot.

Facial skincare products were the first product type. Because the shape of the edge has a certain arc and the diameter is slightly different, it is more suitable to use V-shaped fingers

for one-handed grasping. The driving method was pneumatic, and the diameter of the circumference formed by the fingers was changed by the movement of the air cylinder to achieve the purpose of clamping different diameter pillars. The structure of the mechanical arm is shown in Figure 2a. The second type of goods was frozen and refrigerated. In order to make the grasping effect safe and stable, and to adapt to the shape and size of the grasped goods, the flexible and adaptive grasping method with both hands was adopted. The flexible adaptive gripping method could freely change the finger shape combination of the robot arm, and the gripping process was convenient and flexible. The structure of the robot arm is shown in Figure 2b. The third type of goods was stationery, which is composed of various types of crayons, colored pencils, and oil pastels. At the same time, this category also includes some cuboids with different sizes and light weights. The two-hand pneumatic grip with adjustment function was used. The structure of the robot arm is shown in Figure 2c. The fourth type of products was beverages. Considering that the packaging material of some alcoholic products is glass, if the clamping force is too large, the bottle may be broken. Therefore, a three-jaw chuck with a suction cup was designed for gripping. The structure of a three-jaw (three-finger) robot arm is shown in Figure 2d. The fifth category was laundry products. Because the weight of the grabbed goods range from 110 to 4260 g, the hydraulically driven two-hand grip was adopted, which has a large gripping force and high control accuracy. The structure of the robot arm is shown in Figure 2e. The sixth category was puffed foods, such as biscuits, cakes, puffed foods, etc., which are fragile items. The bionic software robot with suction cups was used to grab the goods to protect the goods from being damaged during the grasping process. Owing to the smooth surface of this type of packaging, vacuum adsorption can be formed smoothly. The structure of the robot arm is shown in Figure 2f. The seventh category was paper products, including pumping paper, wet wipes, toilet paper, facial tissues, sanitary napkins, and sanitary pads. The corresponding gripping method was a two-hand retractable splint clamp with a stepper motor as the gripping power. The structure of the robot arm is shown in Figure 2g. The eighth category of goods was miscellaneous goods, including objects of various shapes, such as towels, coins, water cups, power banks, hangers, shoe brushes, toothbrushes, hooks, and other irregular items. The corresponding gripping method was to use electric-powered bionic manipulators for gripping, and various gripping operations could be performed. The structure of the robotic arm is shown in Figure 2h.

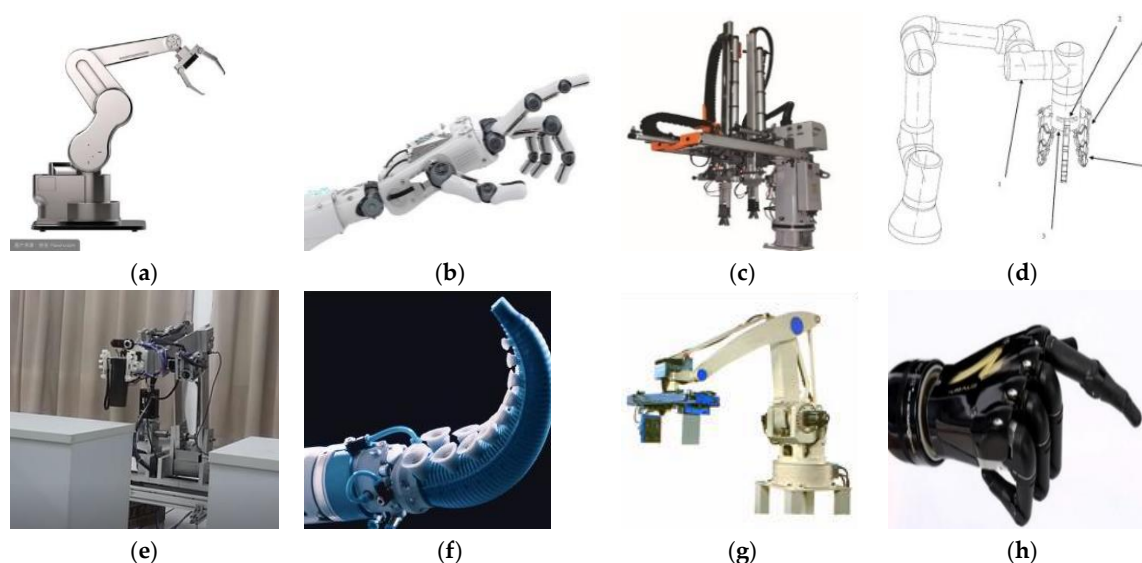


Figure 2. Different grabbing schemes for different types of goods: (a) V-shaped finger grip with pneumatic drive; (b) flexible adaptive grabbing with both hands; (c) adjustable two-hand pneumatic clamp; (d) three-jaw chuck with suction cup; (e) hydraulic drive; (f) bionic software robot grip arm with suction cup; (g) retractable plywood; (h) I-limb robot arm.

According to the matching of the grasping scheme and the type of goods, and the picture data of the goods obtained, the different types of goods could be identified and the corresponding grasping scheme could be matched.

3.3. Recognition Model Construction

A convolutional neural network model with deep learning was established for goods identification. The image recognition of the goods images obtained was carried out to complete the matching of the grabbing types for the goods. The specific model framework is shown in Figure 3.

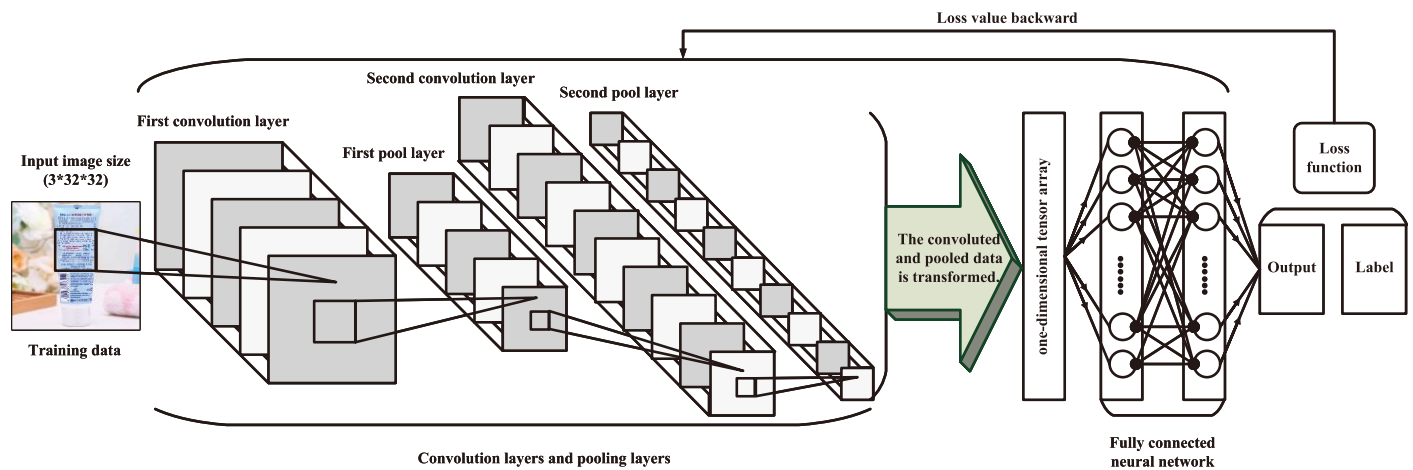


Figure 3. Training process of image recognition and classification based on a convolutional neural network.

In this paper, the neural network structure of the double convolution layer and double pooling layer is adopted, and the convolution and pooling results are transformed into a fully connected neural network, and finally the image classification and recognition results are output. The image data used in training and testing are color images with three channels. The image features are extracted by convolution and a pooling operation for classification and recognition. The training results are compared with the labels, and the parameters of each part of the convolution neural network are modified by the value of the loss function returned, so as to continuously improve the accuracy of the model for image classification and recognition. Through the distribution of all the training set image information into the model for repeated training, the trained network model is finally obtained, and it is applied to the classification and recognition of the test data set to judge the final effect of model training.

3.3.1. Feature Extraction

A four-dimensional tensor $x^l \in \mathbb{R}^{H^l \times W^l \times D^l \times D}$ is used to represent a data set containing multiple pictures. The four dimensions of the data set are: the pixel length of the picture in the data set, the pixel width of the picture, the number of picture channels, and the number of samples in each batch. The input tensor of l -th convolutional layer in the three-dimensional case is $x^l \in \mathbb{R}^{H^l \times W^l \times D^l}$, the triple (i^l, j^l, d^l) is used to indicate that the tensor corresponds to the element at the i^l -th row, the j^l -th column, and the d^l -th channel position, where $0 \leq i^l \leq H^l, 0 \leq j^l \leq W^l, 0 \leq d^l \leq D^l$, and the convolution kernel of this layer is $f^l \in \mathbb{R}^{H^l \times W^l \times D^l}$. In the case of 3D input, the convolution operation actually extends the 2D convolution to all the channels D^l at the corresponding position, and the convolution result at that position is the sum of all HWD^l elements in a convolution process. If there are D convolution kernels such as f^l , the convolution output of $1 \times 1 \times 1 \times D$ dimensions

can be obtained at the same position, and D is the number of channels about the feature x^{l+1} from $l + 1$ -th layer. The formal convolution operation can be expressed as:

$$y_{i^{l+1},j^{l+1},d} = \sum_{i=0}^H \sum_{j=0}^W \sum_{d^l=0}^{D^l} f_{i,j,d^l,d} \times x_{i^{l+1},j^{l+1},d^l}^l, \quad (1)$$

where (i^{l+1}, j^{l+1}) is the position coordinate of the convolution result, and it satisfies the following equations:

$$0 \leq i^{l+1} < H^l - H + 1 = H^{l+1}, \quad (2)$$

$$0 \leq j^{l+1} < W^l - W + 1 = W^{l+1}, \quad (3)$$

where $f_{i,j,d^l,d}$ can be regarded as the learned weight. In addition, a bias term b_d needs to be added to $y_{i^{l+1},j^{l+1},d}$.

The tensor of the first convolution layer input contains four dimensions: the length of the picture (32 pixels), the width of the picture (32 pixels), the number of channels of the picture (3 channels), and the number of samples in each batch (128 pieces). The number of output channels in the first convolution layer is 32, and the size of convolution kernel is 5. In the process of image data convolution, each step jumps once, and the size of 0 added around each channel datum after convolution is 2. The pool size of the first pool layer is 2. The number of input channels of the second convolution layer is 32, and the number of output channels is 32. The size of the second pool layer is also 2.

The maximum pooling aggregation layer is used to transmit the maximum value in the sensing domain to the next layer. The maximum pooling selects the largest element in each sensing field. In the pooling process, the input image size is $3 * 32 * 32$; $2 * 2$ filters are used, and the step size is 2. The pooling kernel of the l -th layer can be expressed as $p^l \in \mathbb{R}^{H \times W \times D^l}$, and the maximum pooling is:

$$y_{i^{l+1},j^{l+1},d} = \max_{0 \leq i \leq H, 0 \leq j \leq W} x_{i^{l+1} \times H + i, j^{l+1} \times W + j, d^l}^l. \quad (4)$$

3.3.2. Mapping Output

In the whole convolution neural network, the full connection layer acts as a classifier (or a reasoner). According to the overall structure of the model design, the input of the fully connected neural network is the output of the convolution neural network after the second pooling operation. On the basis of repeated experiments, a relatively simple fully connected neural network structure is designed, which contains only one hidden layer. Since the input of the fully connected neural network is required to be a one-dimensional tensor, it is necessary to flatten the output results of convolution layer and pool layer. The size of the one-dimensional tensor after flattening is $32 * 8 * 8$. In addition, the output size of the fully connected neural network is consistent with the number of objects to be grasped.

The activation function layer is also called the nonlinear mapping layer, which can increase the expression ability of the entire network. The advantage of the ReLU activation function is that it has no saturation problem and can significantly reduce the problem of gradient disappearance. The ReLU function is actually a piecewise function, defined as follows:

$$\text{rectifier}(x) = \max\{0, x\} = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}. \quad (5)$$

In addition, in order to avoid overfitting, regularization puts indicators that can describe the complex procedures of the model into the loss function, thereby improving the generalization ability of the model. Generally, the complexity of the model is determined by the weights w . The regularization method used in this study is:

$$L = \frac{1}{2} \lambda \|w\|^2. \quad (6)$$

3.3.3. Batch Processing

In order to avoid the disadvantage of not being able to reach the local optimum, the stochastic gradient descent method based on “batch processing” data was used in this study instead of the stochastic gradient descent algorithm. This method is a combination of standard gradient descent and stochastic gradient descent, which takes 64 or 128 training samples as batch processing values, and can obtain more robust gradient information from a batch of samples than a single sample. Therefore, batch stochastic gradient descent is more stable than classical stochastic gradient descent. The batch value set in the model described here is $N = 128$.

Cross entropy represents the distance between the actual output probability and the expected output probability, which means that the smaller the value of the cross entropy, the closer the two probability distributions are. The objective function used in this study is the cross-entropy loss function; the formula follows:

$$Z = -\sum_i y_i \log(p_i), \quad (7)$$

where p_i is the actual output, and:

$$p_i = \frac{\exp(x_i^L)}{\sum_{i=1}^C \exp(x_i^L)}, i = 1, 2, \dots, C. \quad (8)$$

4. Results and Analysis

4.1. Model Training and Parameter Setting under Three Channels

In this study, we used Pytorch as a tool to implement the model. After rotation, there were 120,000 images, 100,000 of them were selected from clear data as training-set data, and the others used as test-set data. The test program for the prediction effect of the model only returns the predicted value periodically. After a series of continuous adjustments of the model parameters, the learning rate was set to $lr = 0.00002$ and the number of training rounds was set to $episodes = 1000$. In the process of training, while using the training-set data, all parameters of the model were saved periodically, while the test model tested the prediction performance on another part of the test set to judge the accuracy of the recognition of image categories and matching grabbing methods.

In the process of model training, we set up each round to calculate the current loss value function; record and display the training error of training-set data; and assess the accuracy of model training, the test error of the test-set data, and the accuracy of model classification. After iterating all the pictures in the test set and the model run for 1000 rounds, the test error of the goods identification model in the test data set was 0.00075, and the classification accuracy of the model was 89.92%.

4.2. Visualization of the Goods Grabbing Model under Three Channels

4.2.1. Analysis of Training Error and Test Error

To ensure the stable convergence of the training error and test error, a lower learning rate was set in this study. As shown in Figure 4, through 1000 rounds of training, it was found that training error and test error decrease rapidly in the early stage of the training process (before 200 rounds), and even the decline speed of the test error was slightly faster than that of the training error. Although the decline of the two errors in the late training process was relatively gentle, from the overall point of view, both of them could reach a lower state in the end. In addition, in the later stage of the training process, the test error was slightly higher than the training error, but the test error did not show an upward trend, so the overfitting in the model was within an acceptable range, and the model training results were effective.

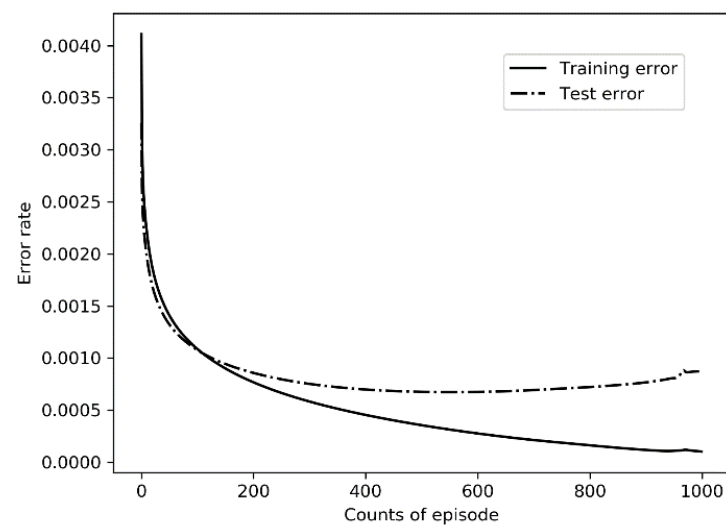


Figure 4. Training error and test error.

4.2.2. Analysis of Training Accuracy and Test Accuracy

As shown in Figure 5, the training accuracy rate and test accuracy rate showed a steady increase in the process of model training, and both of them achieved significant growth in the early stage (before 100 rounds). With the continuous training process, the classification accuracy of the test-set data was slightly lower than that of the training-set data, but it still performed well overall. Through the application of the training results of the model, the image classification was relatively accurate, and the matching between the target goods and the grasping methods was realized, so that the robot arm could select the accurate grasp method to move the goods.

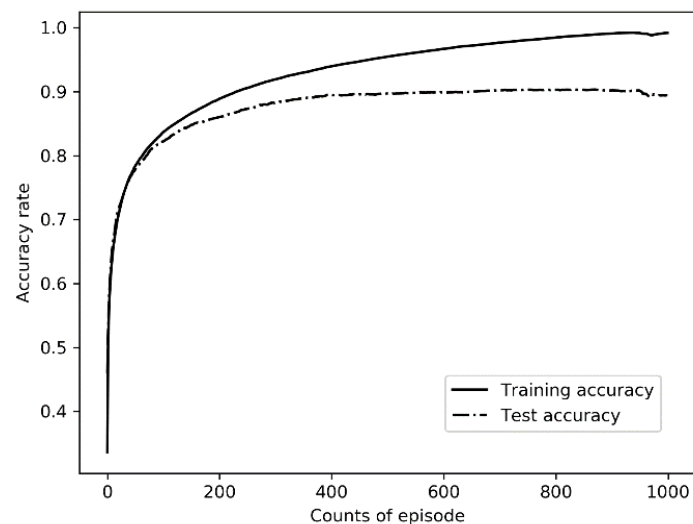


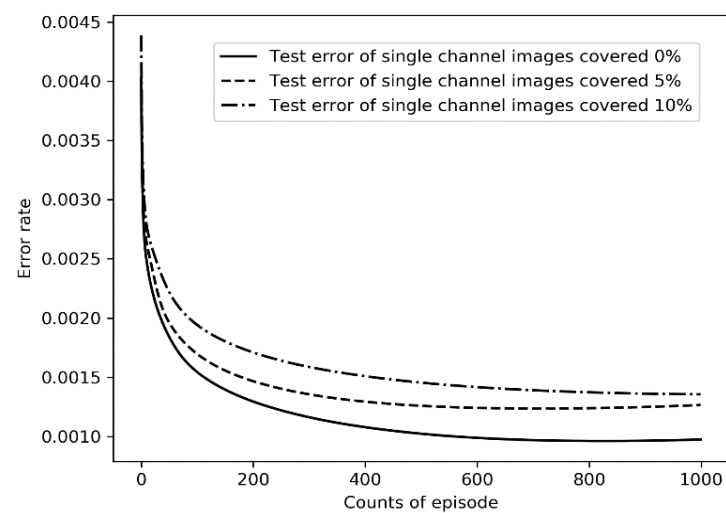
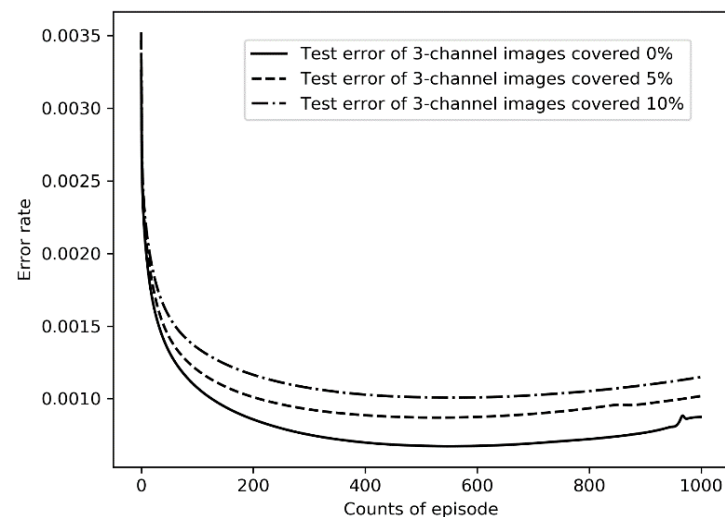
Figure 5. Training accuracy and test accuracy.

4.3. Unusual Conditions

To verify the performance of the system under special conditions such as weak lighting and blurred images, the single-channel images and different rates of cover were selected to simulate these situations, and the training and test were carried out. The results are shown in Figures 6–9 and Table 5.

Table 5. Training and test results for grabbing-type recognition based on images of the goods.

Number of Channels	Covered Rate	Training Error	Test Error	Training Accuracy	Test Accuracy
Single channel	0%	0.000500	0.000962	93.42%	85.49%
	5%	0.000645	0.001236	90.95%	79.47%
	10%	0.001025	0.001356	84.58%	76.49%
Three channels	0%	0.000102	0.000672	99.24%	90.35%
	5%	0.000179	0.000869	98.46%	86.49%
	10%	0.000256	0.001007	97.31%	83.72%

**Figure 6.** Test error of single-channel images.**Figure 7.** Test error of 3-channel images.

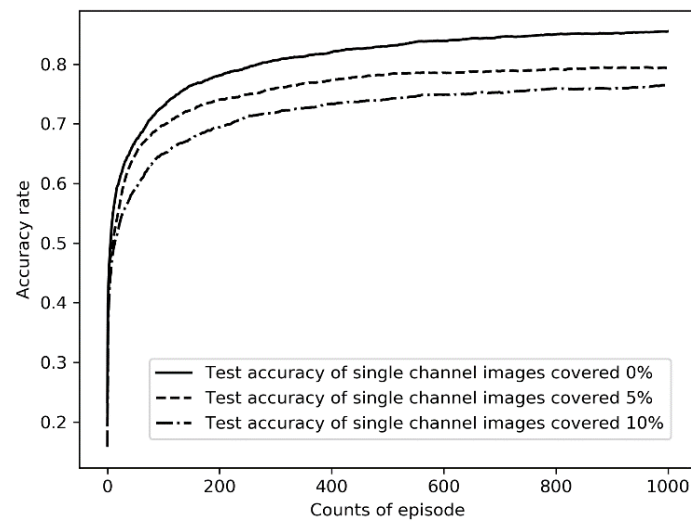


Figure 8. Test accuracy of single-channel images.

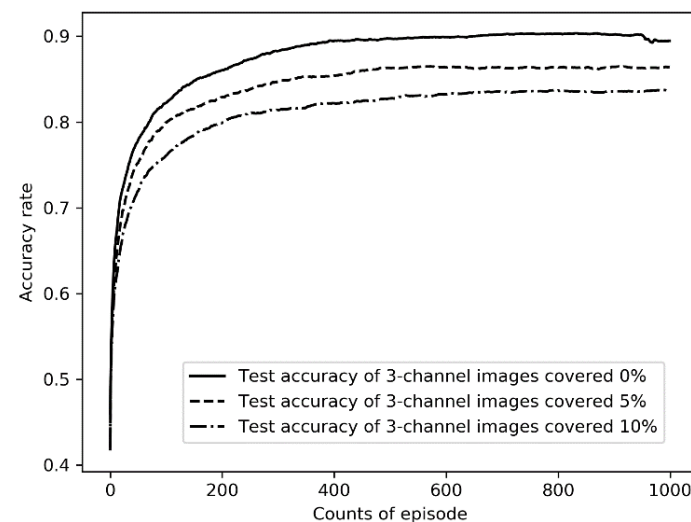


Figure 9. Test accuracy of 3-channel images.

- Three-channel pictures can provide more information from more dimensions, and the training results are significantly better than that of single-channel images. Although it takes a little more time to read the three-channel image data than to read the single-channel image data, CNN used in this study makes the training and recognition speed of the three-channel image basically equal to that of the single-channel images through the operation of the convolution layer.
- In the case of different channel numbers and covered rates, the training error is slightly lower than the test error, and the training accuracy is slightly higher than the test accuracy, indicating that there is a certain overfitting phenomenon. However, with the increase in training rounds, the test error is kept at a low level, the test accuracy rate is maintained at a high level, and there is a trend of continuous improvement, and the phenomenon of overfitting is also within the acceptable range.
- With the increase in the covered rate, although the training accuracy and test accuracy decreased slightly, the decreased range of the training accuracy and test accuracy was far lower than the increased range of the covered rate. Especially in the three-channel image data set, when the covered rate reaches 10%, the recognition accuracy still exceeds 83%. This shows that the CNN model selected in this study can adapt to the complex and changeable data environment and meet the requirements of practical application.

Through the analysis of the above results, it can be seen that the algorithm described in this paper can achieve a very high accurate recognition rate whether the object visual picture is clear and unobstructed or the picture is partially occluded. This also shows that the algorithm we designed has strong adaptability, can cope with different intelligent grasping environments, and successfully completes the task of matching object recognition and manipulator grasping types.

5. Conclusions

The scientific novelty of this study is reflected in the application of COBOT in the field of logistics and warehousing. Through the visual recognition mechanism of a deep neural network, the intelligent classification of objects grasped by a manipulator is realized. It not only expands the theory of intelligent supply-chain and digital logistics in the field of warehouse picking, but also provides a new theoretical research perspective for sustainable development and efficiency improvement of warehouse robots in the field of unmanned picking and warehouse automation. Based on the analysis of the research status of the logistics robot, CNN, and target recognition, this paper focuses on the automatic picking process, using the robot with the image recognition system to identify the type of goods, and further matching the appropriate robot arm for grasping. The identification image data of the goods were obtained from an open e-commerce platform through a web crawler. The driving methods and grabbing types of robot arms were determined based on the characteristics of daily goods and the structures of robot arms. According to the basic framework of CNN, the deep-learning model of intelligent grasping-type recognition was established by setting specific parameters according to specific problems. Finally, the model was trained and its effectiveness was further verified.

In this paper, we apply the CNN model in the field of in-depth learning to the picking activities of warehouse goods, so as to realize the intelligent grab in the picking process and improve the work efficiency of warehouse picking. We use the advantages of a neural network and match different grasping solutions for different kinds of goods. To a certain extent, the research scope of intelligent robots in warehouse picking activities is expanded. In addition, from the practical point of view, the intelligent grabbing solutions of different types and shapes of goods are designed to compensate for the weak link of goods grabbing in picking activities, and promote the realization of the automatic unmanned warehouse in the real sense. In this way, energy conservation and emission reduction can be realized, the concept of green development can be implemented, and the ESG goal in terms of sustainable development can be achieved.

However, the method described in this paper still has some limitations. For example, the objects of visual recognition are static pictures. If video stream information or 3D point-cloud information can be provided, the application space can be further expanded. In addition, future research can also deeply combine visual recognition with manipulator grasping action, from the type matching of manipulator grasping action based on object pictures to the precise design of specific manipulator grasping action, so as to achieve a higher degree of logistics and supply-chain operation intelligence, and promote the sustainable development of logistics and warehousing.

Author Contributions: Conceptualization, L.Z.; Data curation, H.L., J.Z., F.W., K.L. and Z.L.; Formal analysis, J.Z. and F.W.; Funding acquisition, L.Z.; Investigation, H.L.; Project administration, L.Z. and J.Y.; Software, J.Z., F.W. and K.L.; Supervision, J.Y.; Writing—original draft, Z.L.; Writing—review & editing, H.L. and J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the key project of Beijing Social Science Foundation: “Strategic research on improving the service quality of capital logistics based on big data technology”, grant number 18GLA009.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fang, J.Y.; The State Council of China. The Internal Panorama of JD Shanghai “Asia One” Was Exposed for the First Time. Available online: http://www.360doc.com/content/15/1102/15/19476362_510206943.shtml (accessed on 21 June 2022).
2. Khalid, B.; Naumova, E. Digital transformation SCM in View of Covid-19 from Thailand SMEs Perspective. Available online: <https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/pt/covidwho-1472929> (accessed on 21 June 2022).
3. Barykin, S.Y.; Kapustina, I.V.; Sergeev, S.M.; Kalinina, O.V.; Vilken, V.V.; de la Poza, E.; Putikhin, Y.Y.; Volkova, L.V. Developing the physical distribution digital twin model within the trade network. *Acad. Strateg. Manag. J.* **2021**, *20*, 1–18.
4. Lu, X.; Li, B.; Yue, Y.; Li, Q.; Yan, J. Grid R-CNN. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
5. Cui, J.; Zhang, J.; Sun, G.; Zheng, B. Extraction and Research of Crop Feature Points Based on Computer Vision. *Sensors* **2019**, *19*, 2553. [CrossRef] [PubMed]
6. Wang, Y.; Liu, M.; Zheng, P.; Yang, H.; Zou, J. A smart surface inspection system using faster R-CNN in cloud-edge computing environment. *Adv. Eng. Inform.* **2020**, *43*, 101037. [CrossRef]
7. Chen, R.; Wang, M.; Lai, Y. Analysis of the role and robustness of artificial intelligence in commodity image recognition under deep learning neural network. *PLoS ONE* **2020**, *15*, e0235783. [CrossRef] [PubMed]
8. Hong, Q.; Zhang, H.; Wu, G.; Nie, P.; Zhang, C. The Recognition Method of Express Logistics Restricted Goods Based on Deep Convolution Neural Network. In Proceedings of the 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), Xiamen, China, 8–11 May 2020.
9. Dai, Y.; Liu, W.; Li, H.; Liu, L. Efficient Foreign Object Detection between PSDs and Metro Doors via Deep Neural Networks. *IEEE Access* **2020**, *8*, 46723–46734. [CrossRef]
10. Eigenstetter, A.; Takami, M.; Ommer, B. Randomized Max-Margin Compositions for Visual Recognition. In Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
11. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
12. Zhao, Q.; Sheng, T.; Wang, Y.; Tang, Z.; Chen, Y.; Cai, L.; Ling, H. M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Hilton Hawaiian Village, HI, USA, 27 January–1 February 2019.
13. Xie, L.; Liu, Y.; Jin, L.; Xie, Z. DeRPN: Taking a further step toward more general object detection. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
14. Alom, M.Z.; Hasan, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Improved Inception-Residual Convolutional Neural Network for Object Recognition. *Neural Comput. Appl.* **2020**, *32*, 279–293. [CrossRef]
15. Yue, J.; Zhao, W.; Mao, S.; Liu, H. Spectral-spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sens. Lett.* **2015**, *6*, 468–477. [CrossRef]
16. Zhou, W.; Shao, Z.; Diao, C.; Cheng, Q. High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder. *Remote Sens. Lett.* **2015**, *6*, 775–783. [CrossRef]
17. Villalba-Diez, J.; Schmidt, D.; Gevers, R.; Ordieres-Meré, J.; Buchwitz, M.; Wellbrock, W. Deep Learning for Industrial Computer Vision Quality Control in the Printing Industry 4.0. *Sensors* **2019**, *19*, 3987. [CrossRef]
18. Liu, B.; Luo, J.; Huang, H. Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN. *Int. J. Comput. Ass. Rad.* **2020**, *15*, 457–466. [CrossRef]
19. Li, G.; Tang, H.; Sun, Y.; Kong, J.; Jiang, G.; Du, J.; Tao, B.; Xu, S.; Liu, H. Hand gesture recognition based on convolution neural network. *Clust. Comput.* **2019**, *22*, 2719–2729. [CrossRef]
20. Hu, B.; Wang, J. Deep Learning Based Hand Gesture Recognition and UAV Flight Controls. *Int. J. Autom. Comput.* **2020**, *17*, 17–29. [CrossRef]
21. Pigou, L.; Aäron, V.D.O.; Dieleman, S.; Herreweghe, M.V.; Dambre, J. Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. *Int. J. Comput. Vis.* **2018**, *126*, 430–439. [CrossRef]
22. Wan, S.; Goudos, S. Faster R-CNN for Multi-class Fruit Detection using a Robotic Vision System. *Comput. Netw.* **2020**, *168*, 107036. [CrossRef]
23. Pourdarbani, R.; Sabzi, S.; Kalantari, D.; Hernández-Hernández, J.L.; Arribas, J.I. A Computer Vision System Based on Majority-Voting Ensemble Neural Network for the Automatic Classification of Three Chickpea Varieties. *Foods* **2020**, *9*, 113. [CrossRef]
24. Aukkapinyo, K.; Sawangwong, S.; Pooyoi, P.; Kusakunniran, W. Localization and Classification of Rice-grain Images Using Region Proposals-based Convolutional Neural Network. *Int. J. Autom. Comput.* **2020**, *17*, 233–246. [CrossRef]
25. Afrakhteh, S.; Mosavi, M.R.; Khishe, M.; Ayatollahi, A. Accurate Classification of EEG Signals Using Neural Networks Trained by Hybrid Population-physic-based Algorithm. *Int. J. Autom. Comput.* **2020**, *17*, 108–122. [CrossRef]

26. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent models of visual attention. In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
27. Ku, Y.; Yang, J.; Fang, H.; Xiao, W.; Zhuang, J. Optimization of Grasping Efficiency of a Robot Used for Sorting Construction and Demolition Waste. *Int. J. Autom. Comput.* **2020**, *17*, 691–700. [[CrossRef](#)]
28. Zhang, Y.; Cheng, W. Vision-based robot sorting system. In Proceedings of the International Conference on Manufacturing Technology, Materials and Chemical Engineering, Wuhan, China, 14–16 June 2019.
29. Morrison, D.; Corke, P.; Leitner, J. Learning robust, real-time, reactive robotic grasping. *Int. J. Robot. Res.* **2020**, *39*, 183–201. [[CrossRef](#)]
30. Arapi, V.; Zhang, Y.; Averta, G.; Catalano, M.G.; Rus, D.; Santina, C.D.; Bianchi, M. To grasp or not to grasp: An end-to-end deep-learning approach for predicting grasping failures in soft hands. In Proceedings of the 2020 3rd IEEE International Conference on Soft Robotics (RoboSoft), New Haven, CT, USA, 15 May–15 July 2020.
31. Fang, K.; Zhu, Y.; Garg, A.; Kurenkov, A.; Mehta, V.; Li, F.; Savarese, S. Learning task-oriented grasping for tool manipulation from simulated self-supervision. *Int. J. Robot. Res.* **2020**, *39*, 202–216. [[CrossRef](#)]
32. Jiang, P.; Ishihara, Y.; Sugiyama, N.; Oaki, J.; Tokura, S.; Sugahara, A.; Ogawa, A. Depth Image-Based Deep Learning of Grasp Planning for Textureless Planar-Faced Objects in Vision-Guided Robotic Bin-Picking. *Sensors* **2020**, *20*, 706. [[CrossRef](#)] [[PubMed](#)]
33. Coninck, E.D.; Verbelen, T.; Molle, P.V.; Simoens, P.; Dhoedt, B. Learning robots to grasp by demonstration. *Robot. Auton. Syst.* **2020**, *127*, 103474. [[CrossRef](#)]
34. Wang, Z.; Xu, Y.; He, Q.; Fang, Z.; Fu, J. Grasping pose estimation for SCARA robot based on deep learning of point cloud. *Int. J. Adv. Manuf. Technol.* **2020**, *108*, 1217–1231. [[CrossRef](#)]
35. Yang, F.; Gao, X.; Liu, D. Research on Positioning of Robot based on Stereo Vision. In Proceedings of the 2020 4th International Conference on Robotics and Automation Sciences (ICRAS), Wuhan, China, 12–14 June 2020.